Sparse Representation based Face Recognition with Limited Labeled Samples

Vijay Kumar, Anoop Namboodiri, C.V. Jawahar
Center for Visual Information Technology, IIIT Hyderabad, India

***Abstract*—Sparse representations have emerged as a powerful approach for encoding images in a large class of machine recognition problems including face recognition. These methods rely on the use of an over-complete basis set for representing an image. This often assumes the availability of a large number of labeled training images, especially for high dimensional data. In many practical problems, the number of labeled training samples are very limited leading to significant degradations in classification performance. To address the problem of lack of training samples, we propose a semi-supervised algorithm that labels the unlabeled samples through a multi-stage label propagation combined with sparse representation. In this representation, each image is decomposed as a linear combination of its nearest basis images, which has the advantage of both locality and sparsity. Extensive experiments on publicly available face databases show that the results are significantly better compared to state-of-the-art face recognition methods in semi-supervised setting and are on par with fully supervised techniques.**

Keywords*—*Face recognition; semi-supervised learning; sparse representation.

I. Introduction

Face recognition techniques over the years have employed a variety of representations of the face such as Eigenface, Fisherface, Laplacianface, etc. [1]. Each of these techniques, tries to learn a basis or feature space by optimizing an objective function with a specific goal. Eigenface utilizes PCA to maximize the variance of the training images, while Fisherface tries to maximize the separability of the classes in the feature space. In the recent years, ideas from Sparse Representation and Compressed Sensing have been applied to face recognition, where the representation is obtained by an objective function that leads to sparsity [2] in an over-complete basis. The representation over this basis is obtained using l_1-regularized least square approximation. The least square approximation tends to reduce reconstruction error, while l_1-regularization gives rise to sparsity of representation. The test image is thus represented in terms of a few basis images, and is assigned a label of the class that gives minimum reconstruction error. The approach has proved to be very effective and achieves state-of-the-art results.

Motivated from the success of sparse coding, researchers have developed various models for face recognition and other applications [3]. One significant direction is learning dictionaries for reconstructive tasks such as denoising [4], [5] and discriminative dictionaries for classification tasks [6]. Patel *et al*. [7] and Zhang *et al*. [8] explained how the dictionaries can be learned and applied for face recognition using sparse representation. Sparse representation based face recognition methods need a large amount of training data to have an over-complete dictionary, which is essential to obtain a sparse representation. The sparse representation is found to be discriminative [2] as the non-zero weights are mostly at locations corresponding to training images similar to a query image, thus helping in classification. Sparsity, thus is the key for the success of these algorithms, but this has a dependency on the size of the dictionary.

However, in many practical situations we will have limited number of training samples. For example, in automatic face labeling/tagging of photo collections and albums, it is necessary to recognize the faces with limited user-tagged photos for a better user experience. Similarly, in scenarios such as deploying a monitoring system for access control in factories or attendance checks in classrooms, if we can work with a couple of labeled samples that are already collected for other purposes such as ID cards or employee registration, we can avoid the costly phase of manual labeling of faces. Designing an accurate classifier with limited labeled training samples is the focus of this paper.

A possible solution to deal with the lack of labeled samples is to learn a model using both the labeled and unlabeled samples using semi-supervised learning techniques [9]. There are attempts made to deal with lack of training samples. In [10], self-taught learning was proposed based on sparse representation, where a dictionary is learned using unlabeled samples. Labeled samples are coded sparsely over the learned dictionary and a classifier is learned using SVM. Projection based methods, [11], [12], [13] have been proposed for face recognition with few training samples. Roli *et al*. [11], computed the eigenspace using the labeled samples. Unlabeled samples that are closer to the projected mean templates of each class are selected and augmented to the labeled set and the procedure is repeated till all the unlabeled samples are labeled. In Semi-supervised discriminant analysis [13], labeled data was used to infer the discriminative structure of the data while the intrinsic geometric structure of the data is inferred from both labeled and unlabeled samples. Zhao *et al*. [12] used an approach similar to [11] except that feature space is computed using LDA.

In this paper, we propose a graph based multi-stage label propagation algorithm based on sparse representation for semi-supervised face recognition. In each stage, unlabeled samples are labeled using the label propagation algorithm [14] and only highly confident samples are selected. We propose a nearest neighbor based sparse coding (NNSC) algorithm to obtain the graph weights. NNSC is similar to [15], [2] and represents each sample as a linear combination of its nearest neighbors (see Fig. 1). However, it is many times faster than sparse representation based classifier (SRC) with

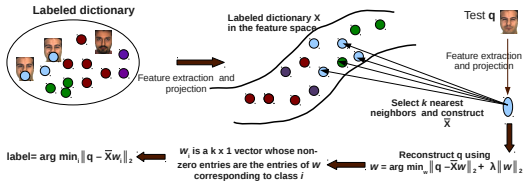


Fig. 1: **Overview of our proposed NNSC algorithm**. Test image is represented as a linear combination of its nearest neighbors and assigned a label of the class which gives minimum reconstruction error.

similar performance. In our representation, a connection is formed only within a neighborhood for each node, while SRC might give connections between far away nodes. We also show an extension of our algorithm for out-of-class samples using a classifier similar to SRC, which tries to minimize the reconstruction error. The query image is assigned the label of the class whose samples minimize the reconstruction error. Experiments on AR and CMU PIE data sets clearly demonstrate the superiority of the approach compared to the existing methods.

II. Sparse Models for Face Recognition

Given an image $x \in \mathbb{R}^d$, along with an over-complete basis or dictionary A $\in \mathbb{R}^{d \times n}$ with *d* elements, n $\gg$ d, the image is represented as a linear combination of "few" elements of the dictionary A. In other words, x is approximated as: $x \approx Aw$, where w is sparse and is computed as:

$$\arg\min_{w} ||w||_0 \quad \text{subject to} \quad x \approx Aw, \tag{1}$$

where $||.||_0$ denotes l_0-pseudo norm and indicates the number of non-zero elements in w. Solving the above l_0 minimization problem is NP-hard and is usually done with greedy methods such as Matching Pursuits (MP) or by l_1-convex relaxation. Dictionary A could be a pre-specified, such as wavelets or training images itself [2], or they could be learned [4], [6] using the following objective function.

$$\arg\min_{A,w} ||w||_0 \quad \text{subject to} \quad x \approx Aw \tag{2}$$

Sparse representation thus could involve two tasks, learning a dictionary, and finding a sparse representation over the dictionary. However, when there are less training samples, it may not be possible to learn a good dictionary that gives a sparse decomposition for the training samples.

A. Supervised and Unsupervised Methods

We have a small set $X_l = \{x_l^{(1)}, x_l^{(2)}, \ldots, x_l^{(m)}\}$ of m labeled examples with their labels $y^{(i)} \in \{1,2,...,C\}$ and a large set of *n* unlabeled examples $X_u = \{x_u^{(1)}, x_u^{(2)}, \ldots, x_u^{(n)}\}$. The subscripts "*l*" and "*u*" indicate labeled and unlabeled images.

In unsupervised sparse coding methods, image is decomposed as a linear combination of a few elements of unlabeled basis. Any image x can be represented sparsely over the dictionary **A**, which is not labeled as shown in Eq. (2). Dictionary A could be pre-specified or learned using training images. These methods are suitable for reconstructive tasks such as denoising [5], [4] or to understand the structure of the data [10].

In supervised sparse coding, label information of the training samples X_l is used to build models that will help classification tasks. In this approach, an image *x* can be coded sparsely over a dictionary **A** that is labeled as shown below.

$$\arg\min_{A,w} ||w||_0 \quad \text{subject to} \quad x \approx \sum_{i=1}^{n} a_i^{(l_i)}.w_i, \tag{3}$$

where $a_i^{(l_i)}$s are basis elements belonging to class l_i. In supervised methods, a discriminative basis can be learned such that samples of each class is best represented by the basis learned for that class [6], [8]. SRC [2] used a similar technique, where a given query image is decomposed over a labeled dictionary (training samples in SRC).

B. Semi-supervised Local Coding

When the number of training samples in X_l are limited, semi-supervised methods that exploits the unlabeled samples X_u to understand the structure of the data can be useful. Unlabeled samples are abundant and easy to collect and better accuracies can be achieved if they are used properly along with the labeled training samples.

Sparse coding in general, requires large number of samples whether in supervised or unsupervised mode. If the images themselves are used as dictionaries [2], then a large number of training images are required to make sure that it is overcomplete to ensure sparsity. Also, learning the dictionary using a few images may not give optimal results. In such cases where the number of samples is less, semi-supervised methods that uses both labeled and unlabeled samples can be employed to improve the classification performance. There are many semi-supervised methods that are used in practice and the reader may refer to [9] for a detailed survey.

III. Semi-supervised Learning for Face Recognition

We will consider an example to understand the scenario of our problem. Assume that the training set consists of a single image for each of the 100 subjects in the AR database [16]. We will use SRC [2], a popular and effective technique to serve as a baseline. The first row of Fig 2 shows a query (Fig. 2(a)) and the first ten training samples (Fig. 2(b)). SRC correctly identifies the given query as belonging to the second class, which is quite similar to the query. However, the query in Fig. 2(c) was not recognized as class two due to the significant expression difference. Now we increase the training set by adding another example per subject that includes some expression variation to the training set (see Fig. 2(d)) and again use SRC to recognize the label of sample in Fig. 2(c). This time SRC correctly identifies the sample. While this is a specific case, the observations holds true across databases as indicated by our experiments given in later sections. In practical situations, when we have limited training samples, supervised methods do not perform well as they are unable to account for the kind of variations encountered in practical situations. Semi-supervised methods can be employed in such cases that uses unlabeled samples to improve the performance of the recognition system.

We now look into the proposed semi-supervised face recognition algorithm based on local sparse coding and its extension to out-of-sample data.

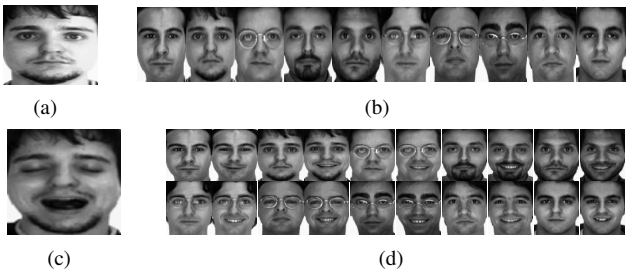


Fig. 2: **Demonstration about the effect of lack of training samples.** SRC identified the (a) first image correctly but failed to identify the (c) second image with (b) small dictionary d_1 (one sample per subject). Using (d) large dictionary (two samples per subject), it identified both images correctly.

A. Transductive learning

Given a set X=$\{x_1, x_2, .., x_l, x_{l+1}..x_n\}$ of training samples. $x_i \in \mathbb{R}^d$, $i = \{1, 2, .., l\}$ are a set of labeled samples belonging to class {1,2,..,c} and remaining samples, x_i, $i = \{l+1, l+2, .., n\}$ are unlabeled. The objective is to predict the labels of the unlabeled samples and subsequently use them to get a better representation of a novel test image to improve the classification performance.

Construct an undirected graph $\langle V, E\rangle$ with similarity matrix W, using both labeled and unlabeled points. Each node in the graph corresponds to a face and the edges E represents similarities between them. Large edge weights w_{ij} indicate that corresponding nodes/faces are very similar. One could use simple k-nearest neighbor method to compute the weights where w_{ij}=1 if x_i is among the k-nearest neighbors of x_j or vice-versa and 0 otherwise. Another option for similarity measure is Gaussian function: w_{ij} = $\exp(-||x_i - x_j||^2/2\sigma^2)$ where σ controls the spread of the Gaussian function.

For face recognition, the above mentioned similarity measures may not be accurate as they are sensitive to illumination and expression variations. Representing a face as a linear combination of training samples is found to be robust to these variations as reported in SRC. Such a method obtains better weights to reveal the relation among different face samples. However, SRC ignores the neighborhood information and thus gives non-zero weights even for the far-away samples in the process of reducing the reconstruction error. Inspired by (SRC) and locality constrained linear coding (LLC), we propose a nearest neighbor based sparse coding (NNSC) that considers both locality and sparsity. In this representation, each sample is represented as a linear combination of its nearest neighbors. We use the NNSC representation to construct the similarity matrix of the graph.

$$\hat{w_i} = \arg\min_{w_i} ||x_i - B_i\ w_i||^2 + \lambda||w_i||_2 \ \text{ s.t } \ \forall_k\ w_{ik} \geq 0 \quad (4)$$

where the columns of B_i are the k-nearest neighbors, $\mathsf{N}(x_i)$ of x_i. We add a positive constraint since graph weights are supposed to be greater than 0. We add a regularization term inspired by CRC [17], which results in a representation that is discriminative. λ is a Lagrangian constant that controls the trade-off between the two terms.

Construct the matrix $W \in \mathbb{R}^{n\times n}$ as:

$$W_{ij} = \begin{cases} \hat{w_i}(p), & \text{if } x_j \in \mathsf{N}(x_i) \\ 0, & \text{otherwise,} \end{cases}$$

where $i, j \in \{1, 2, .., n\}$ and $\hat{w_i}(p)$ denotes the p-th element of vector $\hat{w_i}$. Weights obtained by this method may not be symmetric i.e $w_{ij} \neq w_{ji}$. We make the final weights symmetric with the operation: $w_{ij} = w_{ji} = (w_{ij}$+$w_{ji})/2$. We normalize the weight matrix symmetrically as done in the spectral clustering in order to ensure convergence of label propagation algorithm as shown below.

$$L = D^{-1/2} W D^{-1/2} \ \text{ where } \ D_{ii} = \sum_j W_{ij} \quad (5)$$

Let $F \in \mathbb{R}^{n\times c}$ be a matrix from which the label y_i of a sample x_i, $\{i = 1, 2, .., n\}$ can be obtained as $y_i = \arg\max_j F_{ij}$, where $j = \{1, 2, .., c\}$. For the labeled samples x_i, $\{i = 1, 2, .., l\}$ we define Y_{ij}=1 if y_i=j and 0 otherwise. For unlabeled samples $Y_{ij} = 0$ for all j where j =$\{$1,2,..,c$\}$.

We assume that the label of a sample can be computed as a linear combination of the labels of other samples with the weights being computed through sparse coding of the sample. We propagate the labels of the labeled samples to the unlabeled ones using the constructed graph weights. Using the label propagation framework, we let unlabeled samples receive some amount of label information from its neigbours and retain a part of its initial information at every iteration. The amount of information it receives depends on the corresponding normalized weights.

We begin the iteration with $F(0) = Y$, and for any $t \geq 1$, the labeling matrix F is given by,

$$F(t+1) = \alpha L F(t) + (1-\alpha) Y, \quad (6)$$

where Y is the initial labeling of the samples and α is a parameter that decides the amount of information a sample receives at each iteration. It is well known from the label propagation literature [14] that the above iterative method converges to $F^* = (1-\alpha)(I - \alpha L)^{-1} Y$. The labels of unlabeled samples can be predicted using $y_i = \arg\max_j F^*_{ij}$.

B. Multi-stage label propagation

In the ideal case, each row of F^* contains a single non-zero component corresponding to the true class. The ratio of two largest components in a row in such case will be infinite. However in practice, the ratio will be large only when majority of the samples contributing to the reconstruction belong to a particular class. We can use the ratio to measure how 'confident' is the labeling decision after the convergence of label propagation. We propose a multi-stage label propagation algorithm, where we select only highly confident labelings after the convergence of each stage of the label propagation algorithm. If the ratio of two largest labeling components of a sample i in F^*_{ij} $\forall$ $j = \{1, 2, .., c\}$ exceeds a threshold, the labeling of the sample is considered as confident. Such samples are considered to be labeled for the next stage of label propagation. We will show in our experimental results, how this multi-stage approach gives a significant improvement over single stage label propagation, especially when there are very few labeled samples and the dataset contains large intra-class variations.

Another advantage of this method is that we can reject the samples such as outliers or out-of-database class images, which might otherwise reduce the performance on test images.

(a) (b) (c)

Fig. 3: Examples from (a) Yale (b) AR and (c) CMU PIE databases. (a) and (c) have illumination and lighting variations while (b) has varying expression and illumination.

C. Extension to Out-of-Sample data

The goal of any classification algorithm is to predict the labels of novel test images correctly. To classify a test image, we can use the reconstruction error as the criterion for classification as done in [2], [17]. Given a test image $q \in \mathbb{R}^d$, its representation w over its k nearest neighbors, $B_q = [x_1, x_2, \ldots, x_k]$, is obtained using Eq. (4) (without positive constraint). For each class i, we construct a function $\delta_i : \mathbb{R}^k \to \mathbb{R}^k$ which gets the coefficients associated with the i-th class in B_q.

$$\delta_i^{(j)} = \hat{w}_j \quad \text{if} \quad x_j \in \mathsf{N}(q); \quad j = 1, 2, \ldots, k \tag{7}$$

where the non-zero entries of δ_i correspond to entries belonging to class i from w. Test image is then assigned the label of the class that minimizes the reconstruction error.

$$\text{label}(q) = \arg\min_i ||q - B_q \delta_i||_2 \tag{8}$$

IV. Results and Discussions

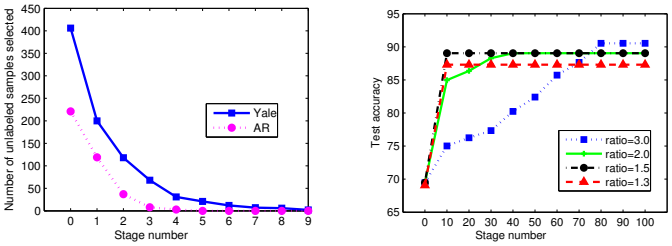
We use Extended Yale B, AR and CMU PIE (shown in Fig. 3) data sets to carry out our experiments. We select the error tolerance, $e = 0.05$ for SRC in all the experiments and choose L1LS [18] l_1-regularized least squares solver to solve the minimization problem in SRC.

A. NNSC-LP Semi-supervised Method

Extended Yale B database [19] consists of 2414 frontal face images of 38 individuals captured under various lighting conditions. Each image is of size 192×168. We resize the image to 80×80. We randomly select half of the images for training (32 images per class) and other half for testing. Few images are shown in Fig. 3(a).

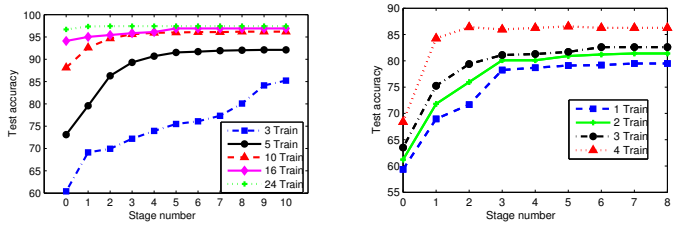
To create a semi-supervised setting, we keep the labels of only few training samples per class (3 to 24) and remove the labels for the rest of training samples. We select the maximum number of stages to 10 and the ratio of two largest labeling component F_{ij}^* (first and second largest labeling component) to $1 : 2.5$. We choose k, that indicates number of nearest neighbors in NNSC to 120. We use both labeled and unlabeled samples to find the feature space using PCA. We select the dimension of eigenface to 504 (to compare with results reported for SRC), $\alpha = 0.9$ and $\lambda = 0.01$.

For different trials, we keep the labels of 3, 5, 10, 16 and 24 training samples in our experiments and labels of rest of the samples are removed. To measure the performance of multi-stage method, test accuracy is calculated after every stage. Fig. 5(a), shows the recognition rates of the test set after every stage for initial labeled samples 3, 5, 10, 16 and 24. It is clear from the figure that for every stage there is an improvement in the accuracy.



(a) (b)

Fig. 4: (a) **Number of unlabeled samples selected after every stage** on Yale and AR database for three labeled examples and threshold ratio=1.5. (b) **Effect of convergence and accuracy** for various values of ratio threshold on Yale database with three labeled samples.



(a) (b)

Fig. 5: (a) **Recognition rates on (a) Extended Yale B database and (b) AR database** vs. NNSC-LP stages for varying labeled samples.

When the number of labeled samples considered are 16 and 24, gain in accuracy is not significant as there are enough labeled samples already available. The significant gain in accuracy can be observed for the trails with 3 and 5 labeled samples where there is almost 20% increase in the recognition rates that indicates the advantage of semi-supervised methods when there are limited training samples.

Fig. 4(a) shows the number of unlabeled samples selected after every stage for Yale and AR database with the ratio of two largest labeling component $F_{ij}^* = 1.5$ and 3 labeled examples per subject. It is clear that, large number of training samples are selected in the first few stages and hence larger gain during these stages. These highly confident samples selected in the first few stages will help in labeling the hard samples in the later stages.

The selection of threshold ratio to select highly confident samples is important for performance. It decides the growth rate of labeled training set at every stage. There is a trade off between accuracy and maximum number of stages as can be seen in Fig. 4(b).

B. Comparison with Other Methods

AR database: We consider a subset of AR face data base [16] consisting of 50 male and 50 female subjects. For each subject there are 14 images with varying expressions and illuminations. Each image is of size 165×120. We convert the images to grayscale and resize to 80×80. Images are taken in two different sessions. We select seven images from session 1 for training and remaining seven images from session 2 for testing. Few images from this database are shown in Fig. 3(b). We select the dimension of eigenspace to 504, $\alpha = 0.9$ and $\lambda = 0.1$. For SRC, we selected a eigenspace dimension that

TABLE I: Recognition rates [%] of various methods on AR database for different number of labeled examples.

Method	1 Train	2 Train	3 Train
SRC [2]	54.4	61.2	65.4
CRC [17]	55.4	62.1	65.5
PCA self-training [11]	62.0	71.0	66.0
LDA self-training [12]	74.5	77.8	80.3
NNSC	59.4	61.6	66.0
NNSC-LP	**79.5**	**81.4**	**82.6**

TABLE II: Recognition rates on CMU PIE database in (mean±std-dev%).

Method	Unlabeled set	Test set
Eigenface	25.3±1.7	25.3±1.6
Laplacianface	56.1±2.3	56.4±2.4
LapSVM [21]	56.5±1.6	56.9±2.6
LapRLS [21]	57.5±1.6	57.9±2.6
SDA [13]	59.0±2.0	59.5±2.7
LDA self-training [12]	84.5±9.5	71.3±6.5
SRC [2]	74.7±1.32	74.9±1.3
CRC [17]	74.9±1.41	75.1±1.32
NNSC	75.0±1.35	74.3±1.3
NNSC-LP	**92.1±1.3**	**92.3±1.5**

maintain a 75% overcomplete dictionary. We set the number of nearest neighbors *k* for NNSC to 100, the ratio of two largest labeling component F^*_{ij} to $1:1.5$ and maximum number of stages to 10. We conduct three trials with 1, 2 and 3 labeled samples. The recognition rates of the test set at various stages is shown in Fig. 5(b).

Table I shows the comparison of our multi-stage NNSC-LP accuracy with other methods. It is clear from the table that, the performance of our NNSC-LP algorithm is superior than other mentioned methods.

C. Single Training Image Face Recognition

CMU PIE consists of 68 subjects with $41,368$ face images with varying illumination, pose, expression and lighting. As reported in [13], [12], we choose a subset of only frontal faces (C27) with only illumination and lighting variations which results in 43 images per subject. The images are cropped to 32×32. We used PCA to reduce the dimension of the image to 504 and *k* is set to 50, $\alpha = 0.9$ and $\lambda = 0.01$. For SRC, dimension is reduced to 50 to have an overcomplete dictionary. We set the maximum stages to 15 and the ratio of two largest labeling component F^*_{ij} to $1:1.5$. For any trial, 30 images are selected for training and remaining 13 are selected for test. Among the 30 training images, only one image is randomly selected and labeled and remaining 29 samples remain unlabeled. The experiment is carried out 20 times and the results are averaged over 20 trials. The results in Table II show superiority of our NNSC-LP algorithm compared to other methods.

D. Effect of parameters: k, λ and α

We observed that the parameters λ and α affected the performance of the results. For all the experiments, we obtained best results with $\alpha = 0.9$ and λ in the range $0.01 - 0.1$. We empirically selected the value of 'k' for NNSC. We found that algorithm is not very sensitive to 'k' and a value of $100 - 200$ seemed to work well in practice.

V. Conclusions

We demonstrate the effect of limited labeled training samples on the accuracy of sparse coding based recognition techniques and how it can be overcome through a semi-supervised approach. The proposed NNSC-LP algorithm accurately labels the unlabeled samples and utilizes them for recognition. NNSC combines the concepts of sparsity and locality. Unlike, SRC, which uses l_1 norm to get the sparse representation, NNSC represents a sample over a few selected neighbors and thus is faster. Experimental results clearly demonstrates the superiority of the proposed method over existing methods when only a few labeled samples are available.

VI. Acknowledgements

This work is partly supported by the MCIT, New Delhi. Vijay Kumar is supported by TCS research fellowship.

References

[1] W.-Y. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, 2003.

[2] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *PAMI*, 2009.

[3] J. Wright, Y. M. Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Y. S. Yan, "Sparse representation for computer vision and pattern recognition," *Proc. of the IEEE*, 2010.

[4] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation." *IEEE Trans. on Signal Processing*, 2006.

[5] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Trans. on IP*, 2008.

[6] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *CVPR*, 2008.

[7] V. M. Patel, T. Wu, S. Biswas, P. J. Phillips, and R. Chellappa, "Dictionary-based face recognition under variable lighting and pose," *IEEE Trans. on IFS*, 2012.

[8] Q. Zhang and B. Li, "Discriminative k-svd for dictionary learning in face recognition," in *Proc. CVPR*, 2010.

[9] X. Zhu, "Semi-supervised learning literature survey," Computer Sciences, University of Wisconsin-Madison, Tech. Rep. 1530, 2005.

[10] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in *Proc. ICML*, 2007.

[11] F. Roli and G. L. Marcialis, "Semi-supervised pca-based face recognition using self training," in *Proc. SSPR/SPR*, 2006.

[12] X. Zhao, N. W. Evans, and J.-L. Dugelay, "Semi-supervised face recognition with LDA self-training," in *Proc. IEEE ICIP*, 2011.

[13] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *Proc. ICCV*, 2007.

[14] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," Carnegie Mellon University, Tech. Rep., 2002.

[15] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *CVPR*, 2010.

[16] A. Martinez and R. Benavente, "The AR face database. CVC technical report #24," June 1998.

[17] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" *ICCV*, 2011.

[18] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale l_1-regularized least squares," *J-STSP*, 2007.

[19] K. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *PAMI*, 2005.

[20] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression (PIE) database," in *Proc. FG*, 2002.

[21] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *JMLR*, 2006.